Targeted Sequential Adaptive Design

Carlo Graziani

1 April 2021

1 Introduction

Suppose we have a space $X \subset \mathbb{R}^D$ and an unknown function $f : X \to Y \subset \mathbb{R}^E$. We take noisy measurements of f, $y_k = f(x_k) + \epsilon_k$, $x_k \in X$, $k = 1, \ldots, N$, where ϵ_k is a zero-mean Gaussian noise term. These measurements give us some idea of the shape of f in the neighborhood of the x_k.

A "classical" adaptive design problem is to sequentially choose sets of points x_k such that the uncertainty of certain functionals of f ("quantities of interest," or "QOI") is minimized. This is the problem of "adaptive experimental design" that was addressed by [2] using techniques of Shannon information theory to characterize uncertainty, and exhibiting a notion of "expected information" from a proposed experiment to optimize the choice of such an experiment. This notion of optimality (also called "D-Optimality") has been usefully coupled to Gaussian process (GP) modeling of the function f, for example in [3].

There is a different question that could be asked in this setting: rather than minimizing the uncertainty of a set of functionals, what if instead we want to find a point $x_T \in X$ such that the function f approaches some desired target value f_T, that is $f(x_T) \approx f_T$? In order to accomplish this goal, we need to make sequential choices of x_k that reduce the uncertainty in $f(\cdot)$, but do so preferentially in regions of X where the value of $f(\cdot)$ approaches f_T.

This is the problem whose solution is set out in this memo. The motivation for the problem comes from the Soderholm group's liquid-liquid extraction (LLE) toy problem, in which the space X is the 3-simplex S_3 of 3-component mixtures, and the space Y is the "custody fraction," also a 3-simplex, representing the fraction of protons associated with each component in the mixture. The idea is in this application is to design experiments that minimize uncertainty while locating a mixture with a desired custody fraction. From the above discussion, however, it should be clear that the problem is considerably more general. As far as I can tell, it is also novel.

The approach adopted here is to assume a GP model on $f(\cdot)$. We will slightly generalize the above notion of measurements using linear functionals of $f(\cdot)$. Then we will define a candidate objective function in terms of the *predictive log-likelihood of* f_T, and show that this function has desirable properties with respect to targeting f_T. We will then show that given a proposed new experiment, it is possible to compute the expectation value of this objective function under the distribution of predicted new data given current data. The objective can thus be optimized simultaneously with respect to the new experimental setting and the choice of x_T.

As we will see, this objective has the desirable property of embodying the "exploration-exploitation" tension — the competition between exploring regions of X where the function $f(\cdot)$ is quite uncertain and investigating regions of X that seem promising for values of x satisfying $f(x) = f_T$. The resulting algorithm is what will be referred to as "targeted sequential adaptive design."

2 Model and Measurements

We will assume a GP model on f. Technically, since f is vector-valued ($f : \mathbb{R}^D \to \mathbb{R}^E$) we require a vector-valued GP (a "VVGP"). This is a straightforward generalization of a GP. A VVGP is characterized by a vector-valued mean

function $\mu_i(x)$, $i = 1, \ldots, E$, and by a matrix-valued covariance function $K_{ii'}(x, x')$ that is positive definite, so that for any vector field $h : \mathbb{R}^D \to \mathbb{R}^E$ we have that $\sum_{ii'} \int dx\, dx'\, h_i(x)\, K_{ii'}(x, x') h_{i'}(x') > 0$. I will not make explicit use of VVGP covariances here, but merely allude to their existence and validity. Examples of their construction and use can be found in [1]. We assume that $f \sim \mathrm{VVGP}(\mu, K)$, for some vector-valued function μ and some valid matrix-valued covariance K.

The choice of covariance K assigns f to a space of functions Γ. A measurement operator on the function $f(\cdot)$ is a linear functional $G : \Gamma \to \mathbb{R}$, yielding the result $G \circ f$. A simple example of such a measurement is a sample at a point $x \in X$, that is, $G_x \circ f = f(x)$. More complex examples extend the scope of the problem. For example, suppose that $f(\cdot)$ is a Fourier transform of some function, so that X is the transformed domain, and $G_t \circ f = \int dx\, e^{itx} f(x)$ is a measurement sample at a space-time point t. Or assume a tomographic setup, where $f(\cdot)$ represents an object to be reconstructed, $\boldsymbol{n}$ is a unit vector in 3 dimensions, $\boldsymbol{y}$ is a detector location, and $G_{\boldsymbol{n},\boldsymbol{y}}$ is the tomographic Radon transform, $G_{\boldsymbol{n},\boldsymbol{y}} \circ f = \int d\lambda\, f(\boldsymbol{y} + \lambda \boldsymbol{n})$. Or G could represent convolution of a telescope image with a point-spread function due to optical diffraction (a "denoising" problem). And so on.

Suppose now we have a set of N measurement operators G_k $k = 1, \ldots, N$, and we write them in a column vector $\boldsymbol{G}$ with components $[\boldsymbol{G}]_k = G_k$. The G_k may be any linear functionals, of the types considered above or of more general types. They will in general contain some parameters (such as sample locations x, tomographic directions $\boldsymbol{n}$ etc. in the above examples) which may be chosen by an experimenter, and which are available for optimization. We will express such parameters θ explicitly, writing (for example) $\mathbf{G}(\theta)$.

An actual observation of $\boldsymbol{G}(\theta)$ yields a vector $\boldsymbol{g} = \boldsymbol{G}(\theta) \circ f + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a zero-mean measurement noise vector with a noise covariance $\left\langle \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \right\rangle \equiv \Sigma$. Since $\boldsymbol{\epsilon}$ represents noise, it is uncorrelated with the VVGP model for f, so that $\langle f \boldsymbol{\epsilon} \rangle = 0$. Because of the VVGP over $f(\cdot)$ and the observation model of $\boldsymbol{g}$, the vector $[f(x), \boldsymbol{g}]^T$ is governed by a joint normal distribution

$$\begin{bmatrix} f(x) \\ \boldsymbol{g} \end{bmatrix} \sim \mathcal{N}\left\{ \begin{bmatrix} \mu(x) \\ \boldsymbol{G}(\theta) \circ \mu \end{bmatrix}, \begin{bmatrix} K(x,x) & K(x,\cdot) \circ \boldsymbol{G}(\theta)^T \\ \boldsymbol{G}(\theta) \circ K(\cdot, x) & \boldsymbol{G}(\theta) \circ K(\cdot,\cdot) \circ \boldsymbol{G}(\theta)^T + \Sigma \end{bmatrix} \right\}. \tag{2.1}$$

Supposing instead that we have two, sequential observations, $\boldsymbol{g}_1 = \boldsymbol{G}(\theta_1) \circ f + \boldsymbol{\epsilon}_1$ and $\boldsymbol{g}_2 = \boldsymbol{G}(\theta_2) \circ f + \boldsymbol{\epsilon}_2$, with $\left\langle \boldsymbol{\epsilon}_1 \boldsymbol{\epsilon}_1^T \right\rangle \equiv \Sigma_1$, $\left\langle \boldsymbol{\epsilon}_2 \boldsymbol{\epsilon}_2^T \right\rangle \equiv \Sigma_2$, and $\langle f \boldsymbol{\epsilon}_1 \rangle = \langle f \boldsymbol{\epsilon}_2 \rangle = 0$. Then, because of the VVGP over $f(\cdot)$ and the observation models of $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$, the vector $\left[f(x), \boldsymbol{g}_1, \boldsymbol{g}_2\right]^T$ is governed by a joint normal distribution,

$$\begin{bmatrix} f(x) \\ \boldsymbol{g}_1 \\ \boldsymbol{g}_2 \end{bmatrix} \sim \mathcal{N}\left\{ \begin{bmatrix} \mu(x) \\ \boldsymbol{G}(\theta_1) \circ \mu \\ \boldsymbol{G}(\theta_2) \circ \mu \end{bmatrix}, \begin{bmatrix} K(x,x) & K(x,\cdot) \circ \boldsymbol{G}(\theta_1)^T & K(x,\cdot) \circ \boldsymbol{G}(\theta_2)^T \\ \boldsymbol{G}(\theta_1) \circ K(\cdot,x) & \boldsymbol{G}(\theta_1) \circ K(\cdot,\cdot) \circ \boldsymbol{G}(\theta_1)^T + \Sigma_1 & \boldsymbol{G}(\theta_1) \circ K(\cdot,\cdot) \circ \boldsymbol{G}(\theta_2)^T \\ \boldsymbol{G}(\theta_2) \circ K(\cdot,x) & \boldsymbol{G}(\theta_2) \circ K(\cdot,\cdot) \circ \boldsymbol{G}(\theta_1)^T & \boldsymbol{G}(\theta_2) \circ K(\cdot,\cdot) \circ \boldsymbol{G}(\theta_2)^T + \Sigma_2 \end{bmatrix} \right\}. \tag{2.2}$$

3 The Objective Function - Preliminary Version

The joint distribution in Equation (2.1) gives rise to the well-known *predictive distribution* for $f(x)$:

$$f(x)|\boldsymbol{g} \sim \mathcal{N}\{m, Q\}, \tag{3.1}$$

$$m(\theta, x) = \mu(x) + \left[K(x,\cdot) \circ \boldsymbol{G}(\theta)^T\right] \left[\boldsymbol{G}(\theta) \circ K(\cdot,\cdot) \circ \boldsymbol{G}(\theta)^T + \Sigma\right]^{-1} \left[\boldsymbol{g} - \boldsymbol{G}(\theta) \circ \mu\right], \tag{3.2}$$

$$Q(\theta, x) = K(x,x) - \left[K(x,\cdot) \circ \boldsymbol{G}(\theta)^T\right] \left[\boldsymbol{G}(\theta) \circ K(\cdot,\cdot) \circ \boldsymbol{G}(\theta)^T + \Sigma\right]^{-1} \left[\boldsymbol{G}(\theta) \circ K(\cdot, x)\right]. \tag{3.3}$$

In consequence of this distribution, we may write the *predictive log-likelihood* of the target vector f_T at the point x:

$$\mathcal{L}(\theta, x) = -\frac{1}{2} \log\det\left(Q(\theta, x)\right) - \frac{1}{2} \left(f_T - m(\theta, x)\right)^T Q(\theta, x)^{-1} \left(f_T - m(\theta, x)\right), \tag{3.4}$$

where an immaterial constant additive term has been dropped. The term proportional to $\log\det Q$ has *not* been dropped – despite its frequent omission when Gaussian "likelihood" expressions are exhibited, it plays an essential role in what follows.

The expression in Equation (3.4) is the log of the predictive probability density at x of the function f having the value f_T, given the observations parametrized by θ. Suppose we were to maximize this expression with respect to x: The result would be the point $x = x_T$ assigning the highest probability to $f = f_T$. This assignment might be made for a couple of reasons. If might be that x_T is in fact a good solution to the problem. Or, it could be that the distribution for f is very vague at x_T, so that the second, data-fit term does not punish f for being "far" from f_T, while the term $\log\det Q$ (which measures the log volume of the "1-sigma" ellipsoid) becomes negative too slowly to have an impact.

In either case, this region of the space now seems interesting to us, and we'd like to seek out new observations that reduce the uncertainty in this neighborhood. If the observation operators are simple sample operators G_x, this means performing a new experiment at the optimal x_T. After updated experimental evidence has become available we will have new functions m and Q, and in general the uncertainty associated with Q will be reduced. We may find that at x_T the narrower Gaussian still ascribes a high probability to f_T. Or, we may discover that x_T was only favored due to lax uncertainty, and the more stringent constraints from later data have favored another region of the space X, which we locate by repeating the optimization of the log-likelihood with the new data.

On reflection, this is not a very satisfactory strategy. The problem is that we can only take one sample at a time this way, which is tedious and inefficient. We ought to be able to simultaneously explore other promising regions of X, at least in the early stages when f is poorly constrained. Also, when the observation operators are not local operators like G_x, it is not clear from the above discussion what should be done to reduce the uncertainty near x_T.

These problems can be addressed by starting from the sequential setup described by Equation (2.2), instead of from the all-in-one setup corresponding to Equation (2.1). The "2" sector in Equation (2.2) corresponds to several proposed experiments to be performed, whereas the "1" sector corresponds to experiments already performed, whose attending data $\boldsymbol{g}_1$ is already in the can. Using Equation (2.2) we express an updated log-likelihood $\mathcal{L}(\theta_1, \theta_2)$, which of course depends on the data in hand $\boldsymbol{g}_1$ and on the hypothetical "latent" data $\boldsymbol{g}_2$. We may average out the $\boldsymbol{g}_2$-dependence of $\mathcal{L}(\theta_1, \theta_2)$ using the distribution $\boldsymbol{g}_2|\boldsymbol{g}_1$ – that is, using what can be known about $\boldsymbol{g}_2$ given our existing measurements (and our model). The resulting expression depends on x and on θ_2, but not on $\boldsymbol{g}_2$, and so we may optimize this averaged expression simultaneously over x and θ_2.

What we will find when this averaging is complete is an objective function that (1) is capable of evaluating many future experiments at a time, rather than a single one at a time, and (2) naturally incorporates an "exploration" uncertainty-reduction imperative that competes with the "exploitation" data-fit imperative, providing a very satisfying solution to the targeted adaptive sequential design problem.

4 Objective Function, No Training Wheels

Here's the computation that was sketched out at the end of the previous section.

4.1 Notation Strip-Down

The notation developed so far for the joint dependencies is essential to the nature of the problem, but risks creating an algebraic train wreck at this point. The normal theory relationships we now require would be heavily burdened if not stripped of the VVGP, the observation operators, the noise, etc.

So, without loss of generality, consider the following joint normal distribution:

$$\begin{bmatrix} f \\ g_1 \\ g_2 \end{bmatrix} \sim \mathcal{N}\left\{ \begin{bmatrix} m \\ m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} C_{ff} & C_{f1} & C_{f2} \\ C_{1f} & C_{11} & C_{12} \\ C_{2f} & C_{21} & C_{22} \end{bmatrix} \right\}, \tag{4.1}$$

which is mapped onto the system in Equation (2.2) by the correspondences

$$
\begin{aligned}
f &\leftrightarrow f(x) && (4.2)\\
g_1 &\leftrightarrow \boldsymbol{g}_1 && (4.3)\\
g_2 &\leftrightarrow \boldsymbol{g}_2 && (4.4)\\
m &\leftrightarrow \mu(x) && (4.5)\\
m_1 &\leftrightarrow \boldsymbol{G}(\theta_1)\circ\mu && (4.6)\\
m_2 &\leftrightarrow \boldsymbol{G}(\theta_2)\circ\mu && (4.7)\\
C_{ff} &\leftrightarrow K(x,x) && (4.8)\\
C_{f1} &\leftrightarrow K(x,\cdot)\circ\boldsymbol{G}(\theta_1)^T \leftrightarrow C_{1f}^T && (4.9)\\
C_{f2} &\leftrightarrow K(x,\cdot)\circ\boldsymbol{G}(\theta_2)^T \leftrightarrow C_{2f}^T && (4.10)\\
C_{11} &\leftrightarrow \boldsymbol{G}(\theta_1)\circ K(\cdot,\cdot)\circ\boldsymbol{G}(\theta_1)^T + \Sigma_1 && (4.11)\\
C_{12} &\leftrightarrow \boldsymbol{G}(\theta_1)\circ K(\cdot,\cdot)\circ\boldsymbol{G}(\theta_2)^T \leftrightarrow C_{21}^T && (4.12)\\
C_{22} &\leftrightarrow \boldsymbol{G}(\theta_2)\circ K(\cdot,\cdot)\circ\boldsymbol{G}(\theta_2)^T + \Sigma_2. && (4.13)
\end{aligned}
$$

The interpretation that we give to the stripped down symbols is similar to that of their more complex cousins: f is to be reconstructed from data g_1 and g_2, which are sequentially obtained — first g_1, then g_2. We proceed by deriving the data-predictive distribution, $g_2|g_1$, and update formulae for the log-likelihood, which we then average according to $g_2|g_1$. In the end, we will use the above correspondences to restore the original interpretation to the calculation.

4.2 Data Predictive Distribution

The data predictive distribution is $g_2|g_1$. This is derivable from Equation (4.1) using the usual normal theory conditioning formula:

$$
\begin{aligned}
g_2|g_1 &\sim \mathcal{N}\left(p^{(2|1)}, Q^{(2|1)}\right) && (4.14)\\
p^{(2|1)} &= m_2 + C_{21}C_{11}^{-1}(g_1 - m_1) && (4.15)\\
Q^{(2|1)} &= C_{22} - C_{21}C_{11}^{-1}C_{12}. && (4.16)
\end{aligned}
$$

4.3 Gaussian Prediction Update Formula

At the stage when only g_1 is known, the predictive distribution for f is

$$
\begin{aligned}
f|g_1 &\sim \mathcal{N}(p^{(f|1)}, Q^{(f|1)}) && (4.17)\\
p^{(f|1)} &= m + C_{f1}C_{11}^{-1}(g_1 - m) && (4.18)\\
Q^{(f|1)} &= C_{ff} - C_{f1}C_{11}^{-1}C_{1f}. && (4.19)
\end{aligned}
$$

Once g_2 has also been ascertained, we have an updated predictive distribution for f:

$$
\begin{aligned}
f|g_1,g_2 &\sim \mathcal{N}\left\{p^{(f|1+2)}, Q^{(f|1+2)}\right\} && (4.20)\\
p^{(f|1+2)} &= m + \begin{bmatrix} C_{f1} & C_{f2}\end{bmatrix}\begin{bmatrix} C_{11} & C_{12}\\ C_{21} & C_{22}\end{bmatrix}^{-1}\begin{bmatrix} g_1 - m_1\\ g_2 - m_2\end{bmatrix} && (4.21)\\
Q^{(f|1+2)} &= C_{ff} - \begin{bmatrix} C_{f1} & C_{f2}\end{bmatrix}\begin{bmatrix} C_{11} & C_{12}\\ C_{21} & C_{22}\end{bmatrix}^{-1}\begin{bmatrix} C_{1f}\\ C_{2f}\end{bmatrix}. && (4.22)
\end{aligned}
$$

It is convenient to express $p^{(f|1+2)}$ and $Q^{(f|1+2)}$ as updates to $p^{(f|1)}$ and $Q^{(f|1)}$, respectively. To do this, we first appeal to the matrix inversion lemma,

$$\begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}^{-1} = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix}, \tag{4.23}$$

with

$$\begin{aligned} I_{22} &\equiv \left(C_{22} - C_{21}C_{11}^{-1}C_{12}\right)^{-1} && (4.24)\\ &= \left(Q^{(2|1)}\right)^{-1} && (4.25)\\ I_{12} &\equiv -C_{11}^{-1}C_{12}I_{22} && (4.26)\\ I_{21} &\equiv -I_{22}C_{21}C_{11}^{-1} && (4.27)\\ I_{11} &\equiv C_{11}^{-1} + C_{11}^{-1}C_{12}I_{22}C_{21}C_{11}^{-1}. && (4.28) \end{aligned}$$

The second line notes the equality with the inverse of Equation (4.16).

Combining Equations (4.21) and (4.23-4.28) we obtain an expression for the mean update:

$$\begin{aligned} p^{(f|1+2)} &= m + \begin{bmatrix} C_{f1}I_{11} + C_{f2}I_{21} & C_{f1}I_{12} + C_{f2}I_{22} \end{bmatrix} \begin{bmatrix} g_1 - m_1 \\ g_2 - m_2 \end{bmatrix} \\ &= m + C_{f1}\left(C_{11}^{-1} + C_{11}^{-1}C_{12}I_{22}C_{21}C_{11}^{-1}\right)(g_1 - m_1) \\ &\quad + C_{f2}\left(-I_{22}C_{21}C_{11}^{-1}\right)(g_1 - m_1) \\ &\quad + \left[C_{f1}\left(-C_{11}^{-1}C_{12}I_{22}\right) + C_{f2}I_{22}\right](g_2 - m_2) \\ &= p^{(f|1)} + \left(C_{f1}C_{11}^{-1}C_{12} - C_{f2}\right)I_{22}C_{21}C_{11}^{-1}(g_1 - m_1) \\ &\quad - \left(C_{f1}C_{11}^{-1}C_{12} - C_{f2}\right)I_{22}(g_2 - m_2) \\ &= p^{(f|1)} + \left(C_{f2} - C_{f1}C_{11}^{-1}C_{12}\right)I_{22}\left[g_2 - m_2 - C_{21}C_{11}^{-1}(g_1 - m_1)\right] \\ &= p^{(f|1)} + \left(C_{f2} - C_{f1}C_{11}^{-1}C_{12}\right)\left(Q^{(2|1)}\right)^{-1}\left(g_2 - p^{(2|1)}\right), \end{aligned} \tag{4.29}$$

where in the last line we have used Equations (4.15) and (4.25).

Similarly, combining Equations (4.22) and (4.23-4.28) we obtain an expression for the covariance update:

$$\begin{aligned} Q^{(f|1+2)} &= C_{ff} - C_{f1}I_{11}C_{1f} - C_{f2}I_{21}C_{1f} - C_{f1}I_{12}C_{2f} - C_{f2}I_{22}C_{2f} \\ &= C_{ff} - C_{f1}C_{11}^{-1}C_{1f} - C_{f1}C_{11}^{-1}C_{12}I_{22}C_{21}C_{11}^{-1}C_{1f} \\ &\quad - C_{f2}\left(-I_{22}C_{21}C_{11}^{-1}\right)C_{1f} - C_{f1}\left(-C_{11}^{-1}C_{12}I_{22}\right)C_{2f} - C_{f2}I_{22}C_{2f} \\ &= Q^{(f|1)} - \left(C_{f1}C_{11}^{-1}C_{12} - C_{f2}\right)I_{22}C_{21}C_{11}^{-1}C_{1f} - \left(C_{f2} - C_{f1}C_{11}^{-1}C_{12}\right)I_{22}C_{2f} \\ &= Q^{(f|1)} - \left(C_{f2} - C_{f1}C_{11}^{-1}C_{12}\right)I_{22}\left(C_{2f} - C_{21}C_{11}^{-1}C_{1f}\right) \\ &= Q^{(f|1)} - \left(C_{f2} - C_{f1}C_{11}^{-1}C_{12}\right)\left(Q^{(2|1)}\right)^{-1}\left(C_{2f} - C_{21}C_{11}^{-1}C_{1f}\right). \end{aligned} \tag{4.30}$$

At this stage, it is probably pretty clear why it was necessary to strip down the notation :-\.

In the next subsection, we will see that the mean update formula, Equation (4.29), is necessary for deriving the required expectation value over the latent future data $\boldsymbol{g}_2$. However, another value of these update formulas is that the associated linear problem ("$\left(Q^{(2|1)}\right)^{-1}$") needs to be solved only in the 2 space of proposed new experiments, given the fixed solution to the linear problem associated with the previous experiments ("C_{11}^{-1}"). This can represent a computational saving over the cost of solving the problems associated with the 2+1 spaces (compare Equations 4.21-4.22).

4.4 Expected Log-Likelihood

The predictive distribution for f with knowledge of g_1 and g_2, evaluated at $f = f_T$, has, according to Equations (4.20-4.22), the log likelihood expression

$$\mathcal{L}(g_2) = -\frac{1}{2}\log\det Q^{(f|1+2)} - \frac{1}{2}\left(f_T - p^{(f|1+2)}\right)^T \left(Q^{(f|1+2)}\right)^{-1}\left(f_T - p^{(f|1+2)}\right), \tag{4.31}$$

where we have indicated a dependency of $\mathcal{L}$ on g_2 that arises through the dependency of $p^{(f|1+2)}$ on g_2. Note that this is the only such dependency in this expression, as the covariance is independent of the data. As adumbrated above, we must now average this expression according to the distribution $g_2|g_1$.

Easily done: Using Equations (4.14-4.16) and (4.29), we have

$$E_{g_2|g_1}\left[p^{(f|1+2)}\right] = p^{(f|1)} \tag{4.32}$$

$$E_{g_2|g_1}\left[\left(p^{(f|1+2)} - p^{(f|1)}\right)\left(p^{(f|1+2)} - p^{(f|1)}\right)^T\right] = \left(C_{f2} - C_{f1}C_{11}^{-1}C_{12}\right)\left(Q^{(2|1)}\right)^{-1}\left(C_{2f} - C_{21}C_{11}^{-1}C_{1f}\right). \tag{4.33}$$

From this it follows that the averaged log-likelihood is

$$\begin{aligned} E_{g_2|g_1}\left[\mathcal{L}(g_2)\right] = & -\frac{1}{2}\log\det Q^{(f|1+2)} - \frac{1}{2}\left(f_T - p^{(f|1)}\right)^T \left(Q^{(f|1+2)}\right)^{-1}\left(f_T - p^{(f|1)}\right) \\ & -\frac{1}{2}\mathrm{Trace}\left\{\left(C_{f2} - C_{f1}C_{11}^{-1}C_{12}\right)\left(Q^{(2|1)}\right)^{-1}\left(C_{2f} - C_{21}C_{11}^{-1}C_{1f}\right)\left(Q^{(f|1+2)}\right)^{-1}\right\}. \end{aligned} \tag{4.34}$$

This is the expression that we will use as an objective for our optimization problem, after restoring the notational superstructure. It has some interesting features. The proposed future "2" experiment has unknown data, but a known predictive covariance $Q^{(f|1+2)}$, and this covariance naturally takes up station in the log-determinant term and in the data-fit term. In default of the "2" data, the role of the mean in the data-fit term is played by the predictive "1" mean $p^{(f|1)}$. If this were all there were, the problem would appear as a slight re-elaboration of D-optimality.

However, there is an additional negative-definite trace term. This term has complicated behavior, but one thing it appears to do is to ensure that new points are not redundant with old points. It does this through the $\left(Q^{(2|1)}\right)^{-1}$ term. In the noise-free case, this term is singular if any of the "2" points reproduces any of the "1" points, for the same reason that the "snake" of a GP predictive distribution has a root-variance "waist" that shrinks to zero as the curve approaches a noiseless training point: the predictive covariance develops a zero eigenvalue as a prediction point approaches a training point. As a consequence, in the noise-free case, when a "2" observation reproduces a "1" observation, the trace term diverges. Therefore, this term causes "1" points to "repel" "2" points, and leads to an exploration incentive that contrasts the exploitation incentive inherent in the data-fit term. In the noisy case the repulsion is softened, but still present: just as the GP snake's waist shrinks to a finite size rather than to zero, so the predictive covariance develops a small (but not zero) eigenvalue. Hence, the strength of the repulsion is regulated by the size of the noise terms. This is in accordance with intuition — very noisy measurements may require multiple nearby samples to get a handle on the shape of f, whereas precise ones don't require much company.

One can similarly see that "2" observations will tend to avoid each other as well, since in the noise-free case they produce redundant rows and columns of $Q^{(2|1)}$, and hence a divergent trace term. So again they have a tendency to disperse that is regulated by the size of the noise terms.

4.5 Piling The Notation Back On

Now we adapt the formula of Equation (4.34) to the Gaussian distribution described by Equations (2.2), which models our targeted adaptive design problem. We appeal to the correspondences in Equations (4.2-4.13), as well as to Equations (4.16), (4.18), (4.19), and (4.30).

The objective function to be optimized over experimental parameters θ_2 and searched over the space X for an $x \in X$ such that $f(x) = f_T$ is $\mathcal{L}(\theta_2, x)$, given by

$$\begin{aligned}\mathcal{L}(\theta_2, x) &= -\frac{1}{2}\log\det Q^{(f|1+2)} - \frac{1}{2}\left(f_T - p^{(f|1)}\right)^T \left(Q^{(f|1+2)}\right)^{-1}\left(f_T - p^{(f|1)}\right) \\ &\quad -\frac{1}{2}\mathrm{Trace}\left\{\left(\mathrm{C_{f2}} - \mathrm{C_{f1}C_{11}^{-1}C_{12}}\right)\left(\mathrm{Q^{(2|1)}}\right)^{-1}\left(\mathrm{C_{2f}} - \mathrm{C_{21}C_{11}^{-1}C_{1f}}\right)\left(\mathrm{Q^{(f|1+2)}}\right)^{-1}\right\}, \end{aligned} \tag{4.35}$$

where

$$C_{ff}(x) = K(x,x) \tag{4.36}$$
$$C_{f1}(x) = K(x,\cdot)\circ \boldsymbol{G}(\theta_1)^T = C_{1f}^T \tag{4.37}$$
$$C_{f2}(\theta_2, x) = K(x,\cdot)\circ \boldsymbol{G}(\theta_2)^T = C_{2f}^T \tag{4.38}$$
$$C_{11} = \boldsymbol{G}(\theta_1)\circ K(\cdot,\cdot)\circ \boldsymbol{G}(\theta_1)^T + \Sigma_1 \tag{4.39}$$
$$C_{12}(\theta_2) = \boldsymbol{G}(\theta_1)\circ K(\cdot,\cdot)\circ \boldsymbol{G}(\theta_2)^T = C_{21}(\theta_2)^T \tag{4.40}$$
$$C_{22}(\theta_2) = \boldsymbol{G}(\theta_2)\circ K(\cdot,\cdot)\circ \boldsymbol{G}(\theta_2)^T + \Sigma_2 \tag{4.41}$$
$$p^{(f|1)} = \mu(x) + C_{f1}C_{11}^{-1}\left[\boldsymbol{g} - \boldsymbol{G}(\theta_1)\circ\mu\right] \tag{4.42}$$
$$Q^{(2|1)}(\theta_2) = C_{22}(\theta_2) - C_{21}(\theta_2)C_{11}^{-1}C_{12}(\theta_2) \tag{4.43}$$
$$Q^{(f|1)}(x) = C_{ff}(x) - C_{f1}(x)C_{11}^{-1}C_{1f}(x) \tag{4.44}$$
$$\begin{aligned}Q^{(f|1+2)}(\theta_2, x) &= Q^{(f|1)}(x) \\ &\quad -\left(C_{f2}(\theta_2, x) - C_{f1}(x)C_{11}^{-1}C_{12}(\theta_2)\right)\left(Q^{(2|1)}(\theta_2)\right)^{-1}\left(C_{2f}(\theta_2, x) - C_{21}(x)C_{11}^{-1}C_{1f}(\theta_2)\right)\end{aligned} \tag{4.45}$$

The solution to the targeteted sequential adaptive design problem is then

$$x_T = \arg\max_x \left\{\max_{\theta_2}\left\{\mathcal{L}(\theta_2, x)\right\}\right\}. \tag{4.46}$$

5 A Bit Of Discussion

As usual with Gaussian/GP problems, this one will threaten to get costly the more data accumulates. Availing oneself of some of the efficiencies that come throught the prediction update formulae will certainly be necessary, especially when searching a batch of N new experimental parameters θ_2.

It is possible that another application of this thing is in manufacturing experiments, such as electrospinning or flamespray chamber experiments, and where the target is a product specification. What I have in mind is that in this case, X is the space of control parameters; Y is a space with some components specifying the product (sizes, plane angles, roughnesses, etc.), and other components specifying observed diagnostic output (e.g. Raman scattering data or output of high-speed photography). The target is the desired product, within some tolerances. Now, only some of the components of the vector $f(x)$ are required to hit the target, the rest are diagnostic data we learn are associated with good or bad product outcomes. I'm have not thought this through, as should be clear from this cursory discussion, but I do believe there are possibilities worth exploring here.

References

[1] Mauricio A Alvarez, Lorenzo Rosasco, and Neil D Lawrence. Kernels for vector-valued functions: A review. *arXiv preprint arXiv:1106.6251*, 2011.

[2] Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, pages 986–1005, 1956.

[3] Jiangjiang Zhang, Weixuan Li, Lingzao Zeng, and Laosheng Wu. An adaptive gaussian process-based method for efficient bayesian experimental design in groundwater contaminant source identification problems. *Water Resources Research*, 52(8):5971–5984, 2016.